June 24, 2025 **VIA ECF**

Hon. Ona T. Wang
U.S. District Judge for the Southern District of New York
Daniel Patrick Moynihan United States Courthouse
500 Pearl Street
New York, NY 10007

Re: *In re OpenAI, Inc. Copyright Infringement Litigation*, No. 1:25-md-03143 (S.D.N.Y.). This Document Relates To: 23-cv-11195

Dear Magistrate Judge Wang:

Pursuant to this Court's June 18, 2025 Order (MDL ECF 217), the OpenAI Defendants and the News Plaintiffs write regarding their efforts to agree on a joint proposed methodology for "sampling [] the 30-day tables of consumer output log data" OpenAI has preserved. MDL ECF 79. The parties met and conferred on June 23, 2025 and were unable to resolve their disputes, which are summarized below per the Court's instruction. MDL ECF 217.[1]

1. **Population samples**

OpenAI's Position:

Samples from 30-Day Tables. The parties agree that the relevant search terms will be run on data from the time period of April 14, 2025 through May 14, 2025 sampled from the 30-day tables OpenAI preserved, and that the sampled data should include (i) conversations OpenAI would retain in the ordinary course, (ii) Temporary Chats, and (iii) conversations that users have attempted to delete manually. News Plaintiffs request three separate sample populations for categories i, ii, and iii. Because the distinction between Temporary Chats and manually deleted conversations is not germane to the task at hand (as defined by the Court, *see* MDL ECF 42), OpenAI proposes two samples, the first consisting of "retained logs" (*i.e.*, category i), and the second consisting of data "marked for deletion" (*i.e.*, categories ii and iii), *id*. ***API data.*** The Court's May 29, 2025 order directing the parties to engage in the present sampling exercise was explicitly limited to "*consumer* output log data." ECF 79 (emphasis added). In the 26 days since that order, News Plaintiffs have never suggested a contrary understanding either in their proposed sampling order, *see* MDL ECF 169, or during the parties' conferrals. At 4 pm ET today, however, News Plaintiffs for the first time demanded that OpenAI expand this sampling exercise to include a single population of data from the API. Setting aside that this is beyond the scope of this exercise,

[1] OpenAI prepared a proposed order reflecting the parties' positions on the sampling methodology and the 4,000+ search terms News Plaintiffs proposed, but News Plaintiffs refused to include it with this filing. OpenAI will file them concurrently to fully comply with notwithstanding the Court's request to submit a "joint proposed sampling methodology." MDL ECF 217. News Plaintiffs explained to OpenAI that it does not make sense to submit a proposed order given the number of outstanding disputes that need to be resolved.

a single sample of API data would serve no apparent purpose without some relevant "Control" against which to compare it. News Plaintiffs' request for API data should be denied. The parties have been separately meeting and conferring about the conversation classifiers for API data per the direction at the May 29 conference. ***Historical ChatGPT consumer data.*** Again departing from the Court's May 29 order, News Plaintiffs demand a sample of *historical* ChatGPT consumer data from April 2023—and, at 4 pm ET today, newly demanded an additional historical sample from January 2024. As a threshold matter, this is technically infeasible: even setting aside the costs of decompression, OpenAI cannot run targeted keyword searches on specific fields across data from 2023 or early 2024. It is also pointless: any historical data samples would consist entirely of data *not* "marked for deletion" and would therefore shed no light on the operative question. ECF 42 at 5. News Plaintiffs' attempts to justify these historical samples have no relation to the operative question and instead relate entirely to merits issues that are fully capable of being evaluated by analyzing the billions ChatGPT consumer conversations OpenAI has already retained and offered for sampling—which necessarily includes data from every relevant time period, including April 2023 and January 2024.

<u>News Plaintiffs' Position</u>:

News Plaintiffs are asking for samples from the following 6 populations (four from April – May 2025 and two from earlier time periods):

(1) ChatGPT data that is ordinarily retained from the time-period of April 14, 2025 through May 14, 2025;

(2) ChatGPT data subject to a user-initiated deletion that OpenAI has been ordered to retain from the time-period of April 14, 2025 through May 14, 2025;

(3) ChatGPT data generated through the "Temporary Chat" feature that OpenAI has been ordered to retain from the time-period of April 14, 2025 through May 14, 2025;

(4) ChatGPT data OpenAI has retained from April 14, 2023 through May 13, 2023, before any of the *MDL* lawsuits were initiated;

(5) ChatGPT data OpenAI has retained from January 1, 2024, through January 31, 2024, the first month after The Times filed its Complaint; and

(6) API data OpenAI has been ordered to retain from the time-period of April 14, 2025 through May 14, 2025.

The selection of these sample populations will accomplish three objectives: (1) show the extent to which output log data to be deleted differs in a significant way from output log data ordinarily retained (populations 1, 2 and 3 above), (2) inform the analysis as to whether OpenAI's post-complaint implementation of News Plaintiff specific blocks make the current output log data materially different from the previous sample populations that have not been irretrievably deleted since post-litigation blocks make it harder to elicit evidence of copying of News Plaintiffs' content (populations 4 and 5 above), and (3) enable sampling of at least some API output logs to inform copying and news-related use cases because OpenAI has deleted virtually all of the previous API output log data (population 6 above). News Plaintiffs have been asking about classifiers with respect to API data since the Court's May 29, 2025 order, but have received no response. OpenAI advised News Plaintiffs for the first time on Monday that they were not preserving all classifier data. These three sampling objectives will help News Plaintiffs ascertain the prejudice that News Plaintiffs have suffered from the deletion of output log data and will aid the Court's Rule 37(e) analysis. News Plaintiffs want to ascertain infringing conduct (for showing prima facie copyright infringement and damages) and news-based use cases (for the fair use analysis and DMCA scienter). News Plaintiffs dispute OpenAI's assertion that keyword searches over sample populations 4 and 5 is "not possible." News Plaintiffs' understanding is that they can be searched after OpenAI decompresses the data. OpenAI should be ordered to decompress and search the output log data from earlier time periods that it chose to make less accessible, including after the litigation was filed. *See* Sedona Conference Commentary on Preservation, Management and Identification of Sources of Information That Are Not Reasonably Accessible, 10 Sedona Conf. J. 281, 282 ("**Guideline 4.** A party should exercise caution when it decides for business reasons to move potentially discoverable information subject to a preservation duty from accessible to less accessible data stores.").

2. Data fields to be searched

<u>OpenAI's Position</u>: From the very beginning of this dispute, both parties' filings and statements have focused on OpenAI's preservation of two categories of "output log data:" the "prompts" users submit, and the "outputs" the models generate. *See, e.g.*, MDL ECF 92 at 10 n.16 (News Plaintiffs' definition). For that reason, after the Court directed OpenAI to preserve "output log data," it placed on hold a 30-day table containing both prompts and outputs rather than "build[ing] something" to preserve its entire "data warehouse." 5/27/2025 Aft. Hearing Tr. 36:20–37:5 (OpenAI explaining this). OpenAI then proposed to run searches over prompts—and now agrees to extend those searches to outputs—in response to the Court's sampling order. ECF 79. News Plaintiffs, however, insist that OpenAI run those searches on "all the logs, records, and data fields ordinarily collected." MDL ECF 171. This is not possible because—as OpenAI explained almost a month ago—the data Plaintiffs seek is not in the "30-day tables." ECF 79 (Order); 5/27/2025 Aft. Hearing Tr. 36:20–37:5. It is also, again, far afield from the purpose of this exercise: this is data users *do not see*, which means it will not reveal whether users choose to delete their conversations in a manner that might be relevant to this litigation. *See* ECF 43 at 2. Even if these data fields could be searched, engaging in this exercise will just confuse the issues.

News Plaintiffs' Position: News Plaintiffs request that all data fields be searched has been limited based on OpenAI's failure to fully comply with the Court's preservation order. Until yesterday, News Plaintiffs assumed that OpenAI had brought itself into compliance with this Court's preservation order, and that it had therefore retained all output log data for ChatGPT and the API platform. Contrary to OpenAI's suggestion, News Plaintiffs never limited the definition of output log data, and have been asking for OpenAI to "preserve everything" pertaining to the output log data. January 22, 2025 Hearing Transcript, p. 36. Consistent with this understanding, News Plaintiffs requested that the searches and keywords they proposed be run over all associated data for the population samples, including prompts, outputs, intermediate prompts and responses from retrieval augmented generation (RAG), and classifiers. However, News Plaintiffs learned for the first time yesterday, during a meet and confer with OpenAI, that despite this Court's order, OpenAI has **not been preserving** the RAG data (including which webpages and content are retrieved as part of RAG) that is present in its normal output logs or all of the classifiers and other metadata about the output log data. This deletion impedes News Plaintiffs' ability to effectively search the output data for, *e.g.*, "news".

3. Proposed searches and keywords

OpenAI's Position: OpenAI agrees to run searches for (a) "paywall" and variants thereof and (b) News Plaintiffs' expanded list of publication names and domains. These searches are designed to evaluate the question posed by the Court's hypothetical—whether those who allegedly use ChatGPT to "evade [News Plaintiffs'] paywalls" are more likely to delete their conversation data. MDL ECF 43 at 2. News Plaintiffs again depart from that hypothetical and demand that OpenAI run over ***4,000 additional search terms***. Those terms include over 3,000 domains unrelated to News Plaintiffs (including, *e.g.*, wikipedia.com and google.com) that were listed on a document OpenAI produced, along with over 1,000 domains from a list of so-called "pink slime" websites, more than half of which are not operative (including, *e.g.*, astrologyquickie.com and chinless.net). None of these have any conceivable relevance to the task at hand. Nor does it make sense to add massively overbroad search terms like "news" or "journalism": to state the obvious, this litigation is about the *specific copyrighted works* asserted by News Plaintiffs, not about *unrelated* works of "journalism" that News Plaintiffs do not own. Adding these terms will not reveal any differences that are material to the question at hand, MDL ECF 42 at 5, and will impose a massive technical burden on OpenAI—whose available search infrastructure was not built to run such a high volume of searches. News Plaintiffs will have ample opportunity to run these kinds of searches themselves when they sample the many billions of conversations that OpenAI has retained. Incorporating them here will completely derail this exercise from its intended narrow purpose.

News Plaintiffs' Position: News Plaintiffs request that OpenAI include searches and keywords related to news content generally and have proposed searches for "news" and related keywords, and specific news-related domains. These searches are directly relevant both to infringement and to the fourth fair use factor (the effect of the use on the market for the copyrighted work). A user query for "What happened with President Trump yesterday," for example, would not trigger OpenAI's proposed keywords, but would presumably be flagged by one of OpenAI's own news-related classifiers. News Plaintiffs are okay with the inclusion of the "paywall" search term, but it seems unlikely that a user seeking news content is likely to include the word in a query. Moreover, if the volume of news-related conversations in the data marked for deletion differs significantly from what is ordinarily maintained, this would be one way in which News Plaintiffs could show the extent and effect of OpenAI's deletions. The list of websites include: (i) known plagiarizers of news content, and (ii) known "pink slime" operations that use AI to create low quality news content.

4. Geographic scope

OpenAI's Position: OpenAI has explained that the Preservation Order is in significant tension with its obligations under foreign privacy laws, *see* MDL ECF 67, and requested that the Order's scope be reasonably limited to data belonging to U.S. users, *see* MDL ECF 127 at 2–3. Subjecting data belonging to non-U.S. users to searches and analysis only exacerbates that tension, which is why OpenAI proposed to limit the relevant sample populations to data belonging to U.S. users. In response to News Plaintiffs recent representations to the Court that they would "amenable to excluding specific jurisdictions" from the scope of the Court's preservation order (MDL ECF Dkt. 33), OpenAI proposed a specific set of jurisdictions where it had specific concerns of the tension imposed by its obligations under foreign privacy laws being most acute. News Plaintiffs, however, refused to even consider OpenAI's proposal unless and until it was reduced to writing.

News Plaintiffs' Position: News Plaintiffs asked OpenAI for a written proposal indicating which countries and laws were at issue, and said News Plaintiffs would consider such a proposal. Moments before this filing, OpenAI requested that conversation data in "the European Economic Area and CH [sic]" and News Plaintiffs will consider this request. OpenAI declined to send one until moments before this filing. News Plaintiffs submit that the Preservation Order should remain in effect for all U.S. and international users, but are amenable to excluding specific jurisdictions to the extent OpenAI raises a specific concern as to a particular foreign privacy law or regulation that prohibits OpenAI's retention of output log data pursuant to the Preservation Order. *See* MDL ECF 169 at 2. Moreover, the international output log data may be highly relevant because many pink-slime AI journalism enterprises are located overseas, as well as many individual users who attempt to evade News Plaintiffs' paywalls or other website blocks. *Id.*

5. Cost shifting

OpenAI's Position: If the Court determines that preservation was unnecessary to advance this litigation—as OpenAI has maintained, *see* MDL ECF 40, 65, 92—then the substantial expenses incurred by OpenAI constitute an undue burden, which justifies cost shifting. *Zubulake v. UBS Warburg LLC*, 217 F.R.D. 309, 318 (S.D.N.Y. 2003).

News Plaintiffs' Position: Cost shifting is not appropriate. "[T]he presumption is that the party possessing information must bear the expense of preserving it for litigation," particularly where (as is the case here) the information is highly relevant to the litigation. *See Treppel v. Biovail Corp.*, 233 F.R.D. 363, 372-373 (S.D.N.Y. 2006). OpenAI's position is that it should be paid for complying with its retention obligations under the Federal Rules of Civil Procedure, and even be paid for the costs of retrieving relevant materials that have been deleted or made less accessible. The Federal Rules, however, require that each party retain electronically stored information relevant to a litigation. *See* Sedona Conference Commentary on Preservation, Management and Identification of Sources of Information That Are Not Reasonably Accessible, 10 Sedona Conf. J. 281, 283 ("Moreover, the Note to Rule 37(e) instructs that parties may not 'exploit the routine operation of an information system to thwart discovery obligations by allowing that operation to continue in order to destroy specific stored information that it is required to preserve.'").

Respectfully submitted,

KEKER, VAN NEST & PETERS LLP
/s/ Edward A. Bayley
Edward A. Bayley
On behalf of OpenAI Defendants

ROTHWELL, FIGG, ERNST & MANBECK, P.C.
/s/ Jennifer B. Misel
Jennifer B. Maisel
On behalf of News Plaintiffs